解　　説　　小特集「能動学習」

5. ロボットの行動獲得のための能動学習

Active Learning for Roboto Behavior Acquisition by Minoru Asada (Emergent Robotics Area, Dept. of Adaptive Machine Systems, Graduate School of Engineering, Osaka University).

浅田　稔[1]

1 大阪大学大学院工学研究科知能・機能創成工学専攻

1 はじめに

能動学習は，環境に対する行動により，訓練データを積極的に得る学習法である．ロボットの行動学習は，実環境に対する実際の物理的な行動による能動学習そのものである．ロボットの行動学習の統一的な枠組は広範なロボット応用の全てをカバーできるほど定式化されていないが，通常の機械学習との関連は，以下のように考えられる [1].

- ロボット学習は意志決定問題を含むので,「能動学習」問題を本質的に含んでいる．
- ロボットの環境は複雑かつ不確実であり，それゆえ多くの試行を必要とするが，これは，実際のロボットのハードウェアで実行するコスト面からは避けたい．
- 多くのロボットシステムは実時間システムであり，意志決定は，実際の時間的拘束を受ける．

ロボット学習法としては，種々のアプローチが提案されているが，最近，反射的かつ適応的な行動を獲得できるロボットの学習法として，強化学習が注目を浴びている [2]．この学習法の最大の特徴は，環境やロボット自身に関する先験的知識をほとんど必要としないところにある．強化学習の基本的な枠組みでは，ロボットと環境はそれぞれ，離散化された時間系列過程で同期した有限状態オートマトンとしてモデル化される．ロボットは，現在の環境の状態を感知し，一つの行動を実行する．状態と行動によって，環境は新しい状態に遷移し，それに応じて報酬をロボットに渡す．これらの相互作用を通して，ロボットは与えられたタスクを遂行する目的行動を学習する．

本稿では，まず強化学習の基本的な枠組を述べた後，実際のロボットへ適用する際の問題点として,「状態空間の構成」,「学習の加速化」,「複雑化への対応」の各問題を指摘し，これらに対して，能動学習の側面から，自律的に状態空間を構成する手法，行動空間を目標に従って構成し学習を効率化する手法，そして多重タスクの遂行で問題になる隠れ状態を能動的に発見する手法を紹介し，現状の問題点とその対処法，今後の展望などを述べる．

2 強化学習の枠組

強化というと，アメリカの行動心理学者スキナーのスキナーボックスが思い出される．鼠を箱の中に入れ，その中にあるレバーを鼠がたまたま押すと，餌がもらえる実験で，一旦レバー押しを憶えると何回もレバーを押し続ける行動をとるそうである (図1参照)．このときレバーを押す行為に正の強化 (餌，報酬，価値などなど) が与えられる．強化学習は，これを確率的動的計画法の枠組で定式化したものである．

鼠は箱の中で，どこにいたり，レバーがどのように見えるかなどの状態 ($s \in \boldsymbol{S}$:状態集合) が分かり，前に進んだり，レバーを押すなどの行動 ($a \in \boldsymbol{A}$:行動集合) をとることができる．このとき，環境は厳密にはマルコフ過程としてモデル化され，現

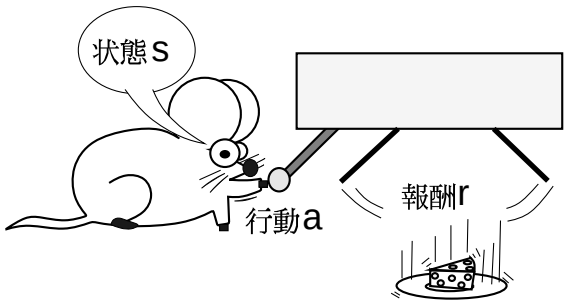


図 1: スキナーの鼠箱

在の状態と鼠がとった行動により確率的に(うまく見えなかったり，脚を滑べらしたりするかもしれないので)次の状態($s' \in \boldsymbol{S}$)遷移する．その結果報酬(r：例えばチーズ)が与えられる．状態遷移が既知であれば通常の動的計画法(以下，DPと略記)の枠組で最適行動が得られるが，未知のとき環境内で試行錯誤しながら，状態遷移と最適行動を推定しなければならない．これが確率的DPとか逐次的DPなどと呼ばれる結縁である．最も良く利用される強化学習法として**Q学習** [3]が有名で，状態sで行動aをとる行動価値関数$Q(s,a)$は，試行錯誤により，次式で更新される．

$$Q(s,a) \Leftarrow (1-\alpha)Q(s,a)+\alpha(r(s,a)+\gamma \max_{a' \in \boldsymbol{A}} Q(s',a')) \quad (1)$$

ここで，s'は，次状態，αは学習率で0と1の間の値をとる．γは，減衰率で，現在の行動が将来に渡ってどれくらい影響を及ぼすかを決めるパラメータで，0と1の間の値をとり，小さい程影響が少ない．行動選択は，学習の収束時間を決める要因の一つで，一旦憶えた成功例を何回も繰り返して上達させるか，別のアプローチを未経験のところから探すかのトレードオフがある．能動学習の観点からは前者が有利であるが，準最適解しか発見できない可能性が高くなる．

3　ロボット学習の課題

強化学習の役割は自律的なエージェントを実現する上で非常に重要であるが，その意義は，より大きく複雑な問題にどの程度適用可能かに依存する．強化学習を始めとするロボット学習を実際のロボットに適用する際の三つの基本的な問題点を以下にまとめる．

・状態・行動空間の構成：従来の強化学習の研究では，多くがコンピュータシミュレーションによるもので，実ロボットへの適用可能性を論議しているものは少ない．ロボットの行動により状態が次状態に遷移する理想的な行動及び状態空間を構成している．例えば，2次元格子状の世界で，ロボットの行動は格子上の上下左右への移動のいずれかであり，状態として格子の座標を対応させるものである[4]．このような状態空間の構成法は，実際のロボットシステムとコンピュータシミュレーションとのギャップを広めている．それぞれの空間は，ロボットが実際感知したり行動できる物理世界と対応すべきと考えられる．しかも，学習が正しく収束できるように構成されねばならない．更に，どのような情報を使えば，タスク遂行に必要かつ十分な状態空間が構成可能であるかを決定する問題も考慮されねばならない．

・報酬関数の構成及び学習の高速化・効率化：Q学習の更新式(1)では，状態遷移毎に即座の報酬が与えることができるが，いい加減な報酬関数では，行動価値関数が多くの極値をもち，学習が収束しない．逆に，収束が保証される詳細な報酬関数が既知である事は，制御則が既知であり，学習を要しない．そこで，単純なタスクでは，最終ゴール状態のみに報酬を与え，それら以外は，ゼロ報酬とする場合が多いが，長い学習時間を必要とする．これを解決する学習の高速化の手法としては，事前にサブタスクに分解する手法[5](分解のための知識を事前に必要とする)，外部から評価者による学習[6](随時，行動の是非を判断する厳密な知識(神様)が必要)，他の学習者の結果を共有する方法[6](経験を共有することによる学習の並列実行による学習時間の短縮化)，簡単なタスクからの実行[7](厳密にやさしさ順で学習を実行すれば，指数オーダーから線形オーダーに短縮．問題は，やさしさの順をどれくらい正確に知っているか?)などが提案されているが，それぞれに一長一短と考えられる．

・複雑化への対応：上記の問題と関連するが，タスクがより複雑になると，サブタスクへの分割だけでは対応できない可能性がある．複数のタスクが互いに独立でなく干渉し，それらを同時に達成し

なければいけない場合が相当する．この場合，1) 複数のタスクに対する行動選択，2) 干渉による隠れ状態の発見，の二つの問題を解決しなければならない．

以下では，最初の二つの問題を考慮した，実ロボットの二つの例を，また最後の問題に対しては，ゴールキーパーを避けながら，ボールをゴールにシュートするロボットの例をもとに，これらの問題点を具体的に明らかにする．

4　状態・行動空間の構成問題

通常の強化学習では，学習が収束するように上手に離散化された状態と行動が定義されている場合が多い．しかしながら一般には，学習が収束可能な状態・行動空間を構成することは容易ではない．

ロボットがタスクを遂行するために必要十分な情報を含む状態空間の構成は，ロボットの行動能力に依存する．また行動空間もロボット自身の知覚能力に依存し，相互に規定しあう．この問題に対し，行動空間を先に固定して，状態空間を構成する手法が提案されている．Chapman and Kaelbring[8] は，TV ゲームの主人公が敵と戦って目的を達成するタスク設定で,「敵を撃つ」,「障害物を回避する」などの構造化された行動をもとに,「敵が部屋にいる」,「ドアが開いている」などの既に抽象化された状態の真偽 (オン/オフ) をビット列とする入力ベクトルを，報酬をもとに分割する手法を提案している．しかしながら，もとの状態が既に抽象化されており，一般的なセンサの実数値の連続空間を対象とする問題には適用できない．

Dubrawski and Reingnier [9] や Kröse and Dam[10] らは，移動ロボットの障害物回避のためのソナー情報の抽象化手法を提案しているが，障害物回避などの局所的かつ反射的なタスクを想定しているので，行動の物理的単位が固定されていても問題は生じにくいが，視覚情報などを利用して遠くにある目標物に到達するタスクなどを想定した場合，同じ物理的動作が，画像上で異なる変化を引き起こし，正しく学習可能な状態と行動を定義することは難しい．例えば，ボールをゴールにシュートするサッカーロボットの例 [11] では，状態数を低減するためにボールやゴールの位置，大きさ，向きを粗く (左中右や大中小など) サンプルした状態空間を用意したが，同じ状態で，同じ行動をとっても異なる状態に遷移することが多くなり，学習が進まない．この問題は**「状態と行動のずれ問題」**と呼ばれ，これに対し，状態が変わるまで，同じ動作を続け，その一連の動作を行動とすることで,「ずれ」が生じないような行動空間を再構築したが，最初に設計者が与えた状態空間がロボットにとって最適である保証はない．

この問題に対し，まず行動選択が一回ですむタスクの例 [12] で，状態・行動空間を構成し，効率的に探索を進める手法をしめそう．基本的な考え方は，成功・失敗例をもとに行動決定木を随時 ID3 で構築し，区間推定法 (Interval Estimation) と呼ばれる手法に基づき，行動を決定する．区間推定法は，$(1-\alpha)$ の信頼性で真の成功確率 P の上下限推定値 $P_{\pm}(\alpha)$ を統計的に求める手法である．最初に各行動は上限値 1 を持ち，ランダムに行動が選択され，試行がある程度経過した時点から，最も高い上限値をもつ行動が通常選択される．区間推定法は，情報と報酬のどちらを獲得する行動を選択するかの問題を解決している．つまり，高い上限値を持つ行動を選択することの二つの理由がある．一つは，情報が少ないために区間が広く，その結果上限値が高い場合，もう一つは，非常に良い行動であるため，信頼性が高く，それ故，区間自体は狭いが，上限値そのものが高い場合である．α は，信頼係数を決定するパラメータであり，小さい値程，情報獲得行動を，大きい程，報酬獲得行動を好む．

図 2にその例をしめす．図では行動 (A)・知覚 (P) 空間が示され，それぞれ 1 次元で表現しているが，一般に多次元にも応用可能である．○×はそれぞれ成功・失敗例を示し，実線で囲まれた箱が ID3 の結果である．今，点線で示された知覚があったとき，模様入りの箱が候補の葉となり，それぞれの葉で区間推定法により，成功確率の上限値を推定し，最も高い上限値を持つ葉 (濃い模様の箱) が選ばれ，行動 A_1 が選択される．

この手法は，ロボットアームで物体を把持する場合の物体へのアプローチ角 (方位角とふ角) を決定する問題に適用された．知覚は物体の距離画像を 2 次超曲面近似した場合の長さ，幅，高さのパラメータである．次元数は少ないが，連続の状態・行動空間を統計的に分割し，少ない試行で成功に

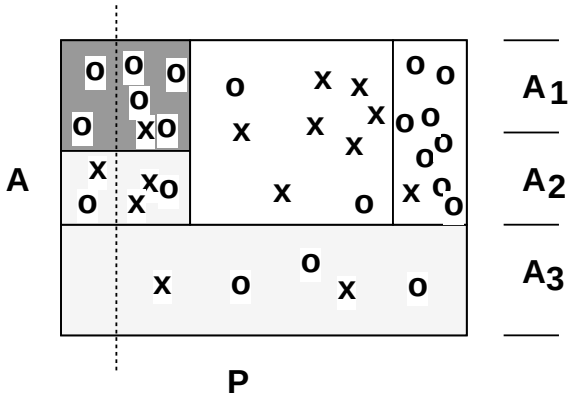
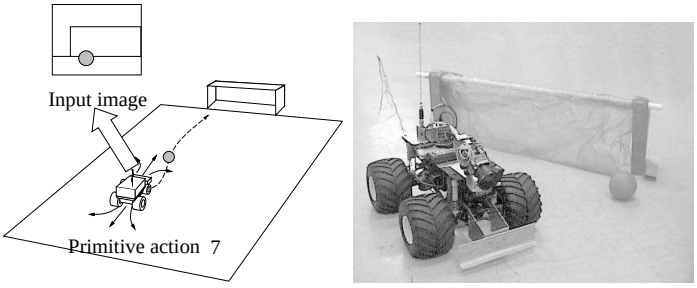


図 2: ID3と区間推定法による行動選択

導く方法である．但し，先にも述べたが，準最適解に留まる可能性があり，αの値の決定が困難と予想される．さらに，一般のタスクを想定した場合，以下の問題があげられる．

1. 図2にも示されているように，この手法では，状態・行動が箱，即ち凸包で表されており，互いに入れ込んだ複雑なタスクへの応用が困難.

2. 一回の行動ではゴールにたどり着けず，行動系列を学習する場合，時間方向への行動分割問題が生じる．



(a) タスク　　(b) 実ロボット

図 3: タスクと実験に用いたロボット

このような例として，視覚移動ロボットがボールをゴールにシュート[1]するタスクを考える [13](図3参照)．画像から得られるのはボールの位置 (x座標) と大きさ，ゴールの位置，大きさ，向きの5次元パラメータであり，行動は二つの独立した左右輪への回転指令である．能動学習の基本的な考え方は，以下である．

1. 5次元の知覚空間は最初2状態 (目標状態とそれ以外) からなる．

2. 行動に関する状態変化を関数近似し，近似による状態変化予測が異なる場合か，ゴール到達に失敗した場合のみ，状態空間を分割し，あらたな関数近似領域を推定する．これにより無駄な探索を軽減できる．

3. 新たに分割された状態の行動価値のみを初期化し，通常の強化学習を適用する (状態数が少ないので学習時間が短い).

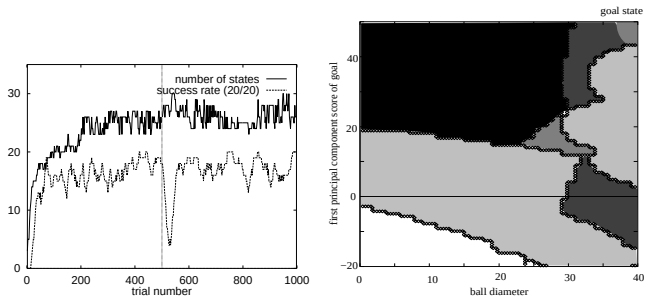
4. 行動選択にランダムネスを付加し，環境の変化に対応する．

図4に実験結果を示す．図左の実線と破線はそれぞれ状態数と過去20試行の成功回数を示している．450回目にボールの大きさを2倍に変更した直後は成功率が下がったが，直ちに持ち直している様子が分かる．図右は分割された5次元の状態空間を2次元 (ボールとゴールの大きさ) に射影したもので，右上がゴール状態を示している．入り組んだ状態が獲得された様子が分かる．状態変化が生じるまで同じモータコマンドを発生させ，その系列を行動と定義することで，行動の時間的分割を行っている．因みに1時間半ほどの実ロボットの学習時間で目的の行動が達成できた．

5　学習の効率化 (複雑化への対応)

強化学習の場合，収束するまでの学習時間は，状態空間のサイズの指数オーダーとなる [6] ので，複雑なタスクに対し，直接強化学習などを利用する事は，ほとんど不可能に近い．一つのアプローチは，サブタスクに分解し，それらを統合することである．Mahadevan and Connel[5] は，ロボットの箱押し作業で，実環境での学習に多大な時間を要するので，作業を事前に「箱の発見」,「箱押

[1]ここでは，単純に押し込む動作で，蹴る行為に相当しないが，サッカーの競技に対応させるために,「シュート」という言葉を用いる．



(a) 成功率と状態数　　(b) 状態空間

図 4: 実験結果

し」,「スタック状態からの回避」の3つに前もって分割し，それぞれに強化学習を適用して，学習時間の短縮化を図った．但し，個々のサブタスクの状態空間が独立で干渉せず，時系列的に実行可能である．また，バンパーセンサー，ソナーなどの近接センサのみを利用しているので，作業の遂行が局所的であり,「箱を指定された場所に運ぶ」などの大局的な目的行動を獲得することには向いていない．

内部ら [14] は，ゴールキーパーを相手にボールをシュートするタスクで，先に獲得したシュート行動 [7] と，別に獲得した回避行動を統合することを考えたが，以下の問題が含まれていた．

- ボールがゴールキーパーに隠されるなどの状態空間が干渉し，知覚見せかけ問題 [15] による隠れ状態が発生する．
- 二つの行動を切替える条件は，状況に依存し容易に決定できない．

そこで，彼らは前者の問題に対し，最尤推定を用いて，隠れ状態を推定・識別した．また後者に対しては，個別に獲得された行動価値関数の単純和を初期値として，隠れ状態近傍を集中的に探索することで，自動的に二つの行動の切替え条件を学習させた．表1にシミュレーション結果を示す．比較のために，回避行動を伴わない「シュート行動」，再学習時の初期値である二つの「行動価値観数の単純和」，ゴールキーパが観測されたとき回避行動をとる「単純切替え」の結果も示す．表から分かるように，再学習した結果が,「成功率」,「衝突までの平均ステップ数(大きい方がよい)」，ゴールまでのステップ数(少ない方がよい)」で最良であった．実ロボットに適用した様子を図5に示す．最初ボールを探すために後退し(第2画面)，右前進してボールを捉え(第4画面)，ゴールキーパーを避けてシュートしている様子が窺える．

表 1: 手法の比較実験

手法	成功率 (%)	衝突までの平均ステップ数	ゴールまでの平均ステップ数
シュート行動のみ	46.7	43.1	286.9
行動価値観数の単純和：再学習の初期値	33.2	77.5	231.2
単純切替え	39.2	98.0	414.4
再学習	46.7	238.1	128.3

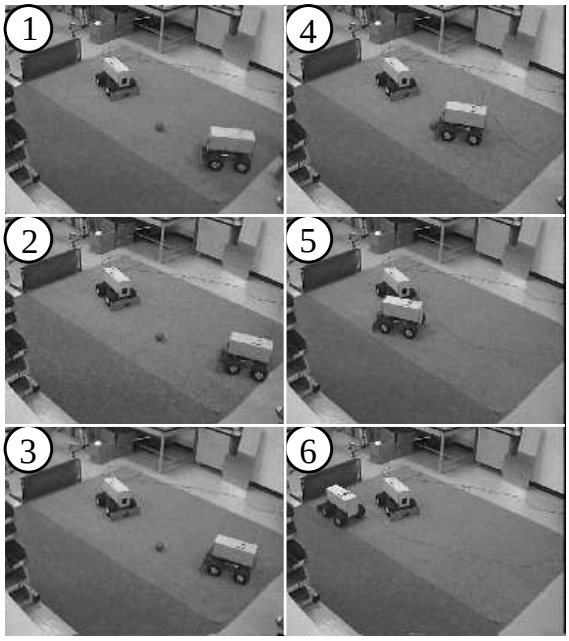


図 5: 実ロボットによる実験の様子

6 おわりに

ロボットの行動獲得手法として，最近注目されている「強化学習」に焦点をあて，実ロボットへの強化学習適用の問題点として,「状態空間構成」,

「学習の加速」,「より複雑なタスクへの対応」について例を用いて説明した.

本文中にも述べたように，示した例は，自由度が少ない移動ロボットであり，行動空間の構成はあまり重要な問題とはなっていない．しかし，多自由度のロボットを制御する場合には，状態空間の構成問題と同様，行動空間の構成問題が重要になり，相互に規定しあう困難な問題となり，これを解決する手法が望まれる.

ロボット学習で対象となっているタスクは，現段階ではまだ，簡単なものであることを否定できない．学習を使わなければ解決困難なタスクをより多く対象とすることで，今後の発展が期待される.

参考文献

[1] J. A. Flanklin, T. M. Mitchell, and S. Thrun. Introduction to robot learning special issue. *Machine Learning*, Vol. 23, pp. 117–119, 1996.

[2] J. H. Connel and S. Mahadevan, editors. *Robot Learning*. Kluwer Academic Publishers, 1993.

[3] C. J. C. H. Watkins. *Learning from delayed rewards"*. PhD thesis, King's College, University of Cambridge, May 1989.

[4] S. Whitehead, J. Karlsson, and J. Tenenberg. "Learning multiple goal behavior via task decomposition and dynamic policy merging". In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 3. Kluwer Academic Publishers, 1993.

[5] J. H. Connel and S. Mahadevan. "Rapid task learning for real robot". In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 5. Kluwer Academic Publishers, 1993.

[6] S. D. Whitehead. "A complexity analysis of cooperative mechanisms in reinforcement learning". In *Proc. AAAI-91*, pp. 607–613, 1991.

[7] 浅田, 野田, 俵積田, 細田. "視覚に基づく強化学習によるロボットの行動獲得". 日本ロボット学会誌, Vol. 13:1, pp. 68–74, 1995.

[8] D. Chapman and L. P. Kaelbling. "Input generalization in delayed reinforcement learning: An alogorithm and performance comparisons". In *Proc. of IJCAI-91*, pp. 726–731, 1991.

[9] A. Dubrawski and P. Reingnier. Learning to categorize perceptual space of a mobile robot using fuzzy-art neural network. In *Proc. of IEEE/RSJ/GI International Conference on Intelligent Robots and Systems 1994 (IROS '94)*, pp. 1272–1277, 1994.

[10] B.J.A. Kröse and J.W.M. Dam. Adaptive state space quantisation for reinforcement learning of collision-free navigation. In *Proc. of IEEE/RSJInternational Conference on Intelligent Robots and Systems 1992 (IROS '92)*, pp. 1327–1332, 1992.

[11] M. Asada, S. Noda, S. Tawaratumida, and K. Hosoda. Purposive behavior acquisition for a real robot by vision-based reinforcement learning. *Machine Learning*, Vol. 23, pp. 279–303, 1996.

[12] Marcos Salganicoff. Active learning for vision-based robot grasping. *Machine Learning*, Vol. 23, pp. 251–278, 1996.

[13] Y. Takahashi, M. Asada, and K. Hosoda. Reasonable performance in less learning time by real robot based on incremental state space segmentation. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems 1996 (IROS96)*, pp. 1518–1524, 1996.

[14] 内部, 浅田, 野田, 細田. 視覚に基づく強化学習による移動ロボットの多重タスク遂行のための協調行動の獲得. 第21回 人工知能基礎論研究会 (SIG-FAI-9403), pp. 25–32, 1995.

[15] S. D. Whitehead and D. H. Ballard. "Active perception and reinforcement learning". In *Proc. of Workshop on Machine Learning-1990*, pp. 179–188, 1990.

浅田 稔(正会員) 1982年大阪大学大学院基礎工学研究科後期課程修了．同年，大阪大学基礎工学部助手．1989年大阪大学工学部助教授．1995年同教授．1997年大阪大学大学院工学研究科知能・機能創成工学専攻教授となり現在に至る．この間，1986年から1年間米国メリーランド大学客員研究員．知能ロボットの研究に従事．1989年，情報処理学会研究賞，1992年，IEEE/RSJ IROS'92

Best Paper Award受賞．1996年日本ロボット学会論文賞受賞．博士（工学）．日本ロボット学会，電子情報通信学会，情報処理学会，人工知能学会，日本機械学会，計測自動制御学会，システム制御情報学会，IEEE R&A, CS, SMC societiesなどの会員